DOCUMENT RESUME

ED 342 317 HE 025 288

AUTHOR Ory, John C.; Bunda, Mary Anne
TITLE There Are Peer Evaluations and There Are Peer Evaluations.
PUB DATE Oct 91
NOTE 18p.; Paper presented at the Annual Meeting of the American Evaluation Association (Chicago, IL, October 1991).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Faculty; Collegiality; Documentation; Evaluation Methods; Higher Education; *Job Performance; *Peer Evaluation; Peer Influence; *Personnel Evaluation; *Portfolios (Background Materials); *Teacher Evaluation

ABSTRACT

This paper discusses the reliability of college faculty peer evaluations. Three evaluation methods are compared: observation; impressionistic; and documentation review. The quality of information provided by each method is examined based on five areas: the purpose of the data being collected; what is being measured; who is responsible for doing the peer evaluation; when the evaluation is conducted; and how the evaluation is conducted. It is noted that evaluation problems still exist, such as judges who lack sufficient knowledge in the candidate's field, too few judges, close collegial relationship, adversarial relationships, lack of anonymity of the judges, and the overall problems inherent in a system that allows decisions to be made without a reasonable rationale. It is suggested that the peer documentation review technique is a critical aspect of peer evaluations, that further research is needed in this area, and that higher education institutions need to supplement student rating information with peer evaluations in annual reviews. Contains 20 references. (GLR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED342317

THERE ARE PEER EVALUATIONS AND THERE ARE PEER EVALUATIONS

John C. Ory
University of Illinois at Urbana-Champaign
Mary Anne Bunda
Western Michigan University

HE 025 288

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mary Anne Bunda

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

THERE ARE PEER EVALUATIONS AND THERE ARE PEER EVALUATIONS

John C. Ory
University of Illinois at Urbana-Champaign
Mary Anne Bunda
Western Michigan University

It is interesting how several generalizations about peer or colleague evaluation of teaching in higher education have become the accepted truths about the subject without clear specifications. It is even more interesting how these beliefs are based on a handful of research-based articles. We cannot read about peer review without being reminded of Centra' 1975 study examining faculty peer observation, Doyle and Crichton's 1978 study comparing student, peer and self-evaluation, Blackburn and Clark's comparison of student and peer ratings, or Batista's 1976 review of the peer evaluation literature. The plea to the historic literature treats the subjects as though there were unanimity of definition, construct and purpose.

While these four articles are aging, their messages are very much alive today. Today's administrator or faculty bargaining unit responsible for developing and implementing a peer review process is warned by Centra (and many others citing Centra's work) that "colleague ratings of teaching effectiveness based primarily on classroom observation would probably not be reliable enough to use in making administrative decisions," primarily because peer ratings are rather "generous" (p.9). Thus, the argument is centered on the issue of reliability rather than validity. Doyle and Crichton shift the focus by reminding that peer ratings are somewhat unrelated to student learning while Blackburn and Clark show that peer ratings are moderately related to

student ratings. Thus, shifting the ground to a validity test, but naming the criterion variables to be either student ratings or student growth. Finally, the Batista literature review of peer evaluation procedures does little but support the notion that "The literature, however, is not very prolific on this topic."

There has developed a general consensus among researchers and practitioners that "peer evaluations" are not very reliable, are marginally valid, are costly and time consuming to conduct, yet are a necessary component in a comprehensive faculty evaluation system. These beliefs have been formed during a time where institutions of higher education are being asked to demonstrate the work that they perform for society. "Trust us," will not suffice as evidence for institutional effectiveness. The professorate is itself under scrutiny. (Cahn, 1986)

Evidence of effectiveness is becoming increasing important in higher education. But what exactly is the evidence we gather when we use peer evaluation?

The most often cited research studies address peer evaluation as though there is only one form. They all use the terms peer or colleague evaluation synonymously in their work. However, the manner in which peer evaluation was performed differed greatly across studies. Centra's colleague evaluation had faculty observing and rating the teaching performance of other faculty. In contrast, Doyle and Crichton's peer evaluation asked faculty to rate "their colleagues' probable classroom presentation by generalizing from such routine experiences as faculty behavior at faculty meetings, colloquia, and social gatherings" (p. 816). Blackburn and Clark asked peer reviewers to make global impressionistic, quality ratings of each of their departmental colleagues.

Most of the accepted beliefs about 'peer evaluations' are based on not one but two types of peer review -- direct observations in classrooms or ratings based on impressions of teaching-types of behavior. Peer observations of teaching are most often thought of when someone speaks of peer , perhaps because the K-12 system has embraced observation in models of clinical supervision. Complaints about poor rater reliability, inadequate sampling and high activity costs are associated with peer visits to the classroom. Slightly more favorable impressions about peer evaluations come from the research studies requiring faculty to rate faculty based on general impressions formed about their colleagues over time. These studies report slightly higher rater reliability and relatedness to outside criteria, such as student ratings than do peer observation studies.

Obviously, there are some serious concerns about both types of peer evaluation procedure. Are faculty willing to spend time observing more than one colleague more than once? How reliable are the ratings of untrained observers? Can we train peer observers to improved the inter-rater reliability? What is being measured? How is validity tested? What criteria are used to make a general impression rating? How confident are faculty in judging their colleagues without the benefit of specific documentation?

In sum, today's general impressions about peer evaluation are based on few research studies employing two distinctive and troublesome methodologies to attain peer ratings of two very different constructs. Do these studies and methodologies represent all there is to say about peer evaluation? Fortunately, not. There is a third methodology for conducting peer evaluations of teaching with a third construct definition of teaching that has received considerable citation but little research attention. The lack of research

involving this methodology has undoubtedly kept it from influencing our common beliefs about peer review.

In her 1981 chapter in the Handbook of Teaching Evaluation, (Millman, 1981) Grace French-Lazovick recommended peer review of teaching documentation that would follow similar procedures used in promotion and tenure reviews of research and service. The teaching documentation would include among other items, instructional materials and evaluation instruments, histories of teaching assignments, personal statements of teaching competence, and description of self-evaluation procedures. Teaching in this case is broadly defined as more than platform or classroom behaviors, but includes judgments of the content of the instructional materials as well. In this case, student feedback data can be used to document classroom behaviors, while the documents can be used to judge the sophistication of the content and the challenge of the exercises designed for students. The author appeared to be ahead of her time in her useful description of peers evaluating what today are referred to as "Teacher Portfolios."

French-Lazovick's chapter is often cited as a useful guide to conducting peer review of teaching. However, the "folklore" of its validity, reliability and utility has yet to develop. While there have been a few encouraging studies of the reliability of peer reviews of research, teaching and service documentation (Spaights and Bridges, 1986; Root, 1987), the authors have been unable to locate studies correlating documentation-based ratings with peer observation or impression ratings, or with student learning, or with student ratings of instruction. This is not to say that research has not been conducted with some of these measures of teaching in an attempt to discover the relationship

between productivity in teaching and productivity in research. (Kremer, 1991; Feldman, 1987) Despite this paucity of research we believe that a comparison of the three peer evaluation methods -- observation, impressionistic, and documentation review -- would further use and study of the documentation review strategy. This comparison of methodologies and focus is offered in the following section.

THE WHO, WHAT, WHEN, HOW AND WHYs of PEER EVALUATION

Each peer evaluation methodology is only as good as the information it provides. Quality evaluation information should be credible, reliable, valid, fair and useful for the intended purpose. One way to compare the quality of the information provided by the different methods is to address the who, what, when, how and why questions of technique. For what purpose is the data being collected? Who does the peer evaluation? What is being measured or evaluated? When is the evaluation conducted? How is the evaluation conducted?

WHY? All three evaluation methods could be used for summative evaluation purposes, i.e., making personnel decisions, albeit one method may be better than another for reasons of credibility, reliability and validity. Faculty development offices have long been using observation or video tapes of teaching as a method of consultation for formative evaluation. If evaluators were the primary force in the introduction of peer information, the argument for peer data would probably be justified by the standards of credibility, reliability and validity. The issue of validity rests on the qualities of the object

measured, credibility rests on the source of the information, while only reliability rests completely with the technique involved. Consequently, evaluators would decide exactly what they wanted to measure, what the best source of that information was, and then proceed to develop a technique to gather reliable data.

Evaluators, however, are not leading the move to gather peer information on instruction. Deciding to involve peers in the evaluation of teaching in any form has at the root not the need evaluators see for triangulation of information, i.e., the supplement of the student rating forms data with another source for validation. Rather, the impetus for the use of faculty ratings of any aspect of teaching comes from a need to improve the status of teaching as an important component of professorship. Research universities are systematically changing the emphasis for tenure to teaching (Grassmuck, 1990). The new focus on teaching quality in higher education has lead departments to involve faculty collecting data and in making explicit judgments of quality rather than using data solely provided by students. The need for a faculty voice in the judgment of teaching was stronger than the need to clearly articulate the meaning of "teaching." Thus, the method of faculty involvement either relies on some common-sense notions of gathering judgments about teaching behaviors or on the literature investigating classroom observation developed in the elementary and secondary school literature where student ratings of teaching are completely absent; where there is a history of research on the observation of classroom behaviors (Rosenshine & Furst, 1973; Medley & Mitzel, 1963). The purpose of the procedure, along with the other data available, should lead to the selection of a methodology and definition of teaching.

For instance, if the data are being collected for a summative personnel decision, clearly the committee will have a host of information available to it for determination of classroom behaviors from the student perspective. The need of the committee is for data which will support some claim of rigor. Observation and document review could also serve formative purposes as both methods can be used to provide suggestions for improving one's teaching. However, the impressionistic method provides a holistic rating only, and fails to identify ways to raise a low rating.

WHO? Who is observing, rating, or reviewing teaching documentation? When the answer to this question can dramatically affect the outcome of the evaluation we believe there are some quality concerns regarding the evaluation method in question. There seems to be some agreement among researchers and practitioners that all peer reviewers, regardless of method, need to be knowledgeable in the content of the faculty members being reviewed and thee departmental, college, or university context of the courses they teach (required/elective, introductory/intermediate or advanced). Presumable the need for context understanding is justified by the need for raters to make judgments of the quality of questions or the quality of examples. Observation of classrooms are not merely an objective recording of the behaviors exhibited by the professor, but rather an informed expert judgment of the material presented. This condition is only relevant when the "**what**" of peer review is consider.

Research suggests additional evaluator requirements for some methods over others. Peer observers should be trained in observation techniques (Centra, 1975) or have experience in observing and offering feedback to teachers. Training in observation

of classroom behaviors is only relevant when the observation guide focuses on teaching behaviors. In most cases this training is not provided for reasons of lack of faculty time, interest or funding. Research by Kremer (1990) indicates the need for impressionistic raters to be very knowledgeable about the faculty being evaluated to ensure reliable ratings. In large departments such familiarity may not be possible Furthermore, high levels of familiarity may not be desirable for fair and unbiased review (French-Lazovick, 1981).

Peers conducting documentation review should be experienced in evaluating dossiers of work and work activities. Many associate and full professors acquire this experience by sitting on departmental executive committees responsible for awarding annual salary increments, or departmental/college/university promotion and tenure review committees.

Therefore, a common element of all of the methods and foci is that the peer involved in the observation is a departmental peer. None of the literature has suggested that the observation can be conducted by a naive recorder of the classroom. A second element of the "**who**" question is how many people should be involved in the rating. Concern for the reliability of the ratings not only requires multiple visits to the classroom in the case of pure observation, but also the use of multiple raters. The use of multiple raters in any of the situation which requires professional judgment is not foreign to the research review traditions of three to five editorial referees.

What? The answer to this question is clearly linked not only to the purpose of the

evaluation, but also to the other data available to decision-making committees. But for the purposes of this discussion, we will consider each of the methods in isolation. The focus or object of the peer evaluation differs significantly across evaluation methods. The least focused of the three is the impressionistic or holistic rating. Anything and everything about teaching can possibly go into making an impressionistic rating. In comparing the other two methods, classroom observation is more limited in its focus as its intent is to evaluate the "classroom teaching performance" of the faculty member. Teaching style, student rapport, and content are often evaluated by peer observers. Documentation review evaluates a broader set of teaching components, including course content, course examinations, advising, teaching philosophy, self-evaluations, course materials and student evaluations of teaching. Clearly, these two operationally define teaching in two dramatically different ways.

Recent arguments by Scriven (1988) raised serious questions about the validity (let alone our capability) of evaluating teachers on the basis of style as is done by many peer observers. He argues for "style-free" evaluation (1991) that refrains from looking for teaching style indicators such as "use of eye contact, advance organizers, enthusiasm, time-on-task" (p. 2). Scriven acknowledge the fact that research on teaching suggest such style indicators tend to distinguish successful from unsuccessful teachers to a statistically significant degree. However, he argues that there is no single, widely accepted definition of good teaching and that

> no indicators whose connection with merit is merely statistical can be legitimately used in personnel evaluation. On this view one can only use the actual performance of the individual being evaluated, on the job-required tasks (the duties of the job). Appealing to anything else is said to be a risk

> of 'guilt by association.' Moreover, it is said that if an evaluation process includes any reference to just one of these style factors -- either in a checklist or amongst the factors that observers consider -- this contaminates the whole process, just as a single question about a candidates's private live contaminates a whole application form. (p.2)

Scriven's argument for not including any of the behaviors he calls 'style' would eliminate many of the items currently used on student rating forms and invalidate many style observations made by peer observers. His point about the definition of an accepted definition of teaching, however, is critically important -- no matter what variables, methods, or sources are to be used in the evaluation.

WHEN? The timing of the peer evaluation is not as critical with the impressionistic and documentation review methods as it is with observation. How often are classes visited? Are they visited at the beginning, middle or end of the semester? Are they conducted when the major presentation is a lecture, classroom discussion or lab exercise? Obviously, the answers to these questions can dramatically impact the evaluation results. The instrumentation used in observation, then, must not have only traditional correlational validity, but some attempt must be made to secure generalizability evidence.

HOW? The impressionistic raters are asked to make holistic ratings of their peers based on past experiences with and knowledge of their colleagues. Observations are usually conducted through an individual visit(s) to the colleagues's class. Documentation reviews involve individual reviews of materials that are later discussed in a group setting.

These variations in methodology may account for differences in the way faculty perceive the credibility of the results.

The impressionistic methodology of peer review was a handy technique for collecting "quick and dirty" peer ratings for research studies comparing peer evaluation ratings with rating derived from other techniques or with external criteria such as student learning. However, the subjective process of rating faculty based on general impressions of teaching quality fails to provide much credibility to the use of impressionistic ratings in personnel decision making.

In comparing the credibility of the other two peer evaluation methods faculty may give the advantage to documentation review. Faculty are probably more confident in reviewing written documentation than in observing the teaching performance of a colleague. This confidence most likely comes form "long training in the evaluation of evidence [that] enables them to weigh what is revealed through documentation"(French-Lazovick, 1981, p. 75). It may also come from experience on promotion and tenure or annual salary review committee that follow a similar methodology for judging research and service, but don't often force an explicit judgment. Faculty may also give greater credibility to a methodology that, as described by French-Lazovick, uses a group consensus rather than relying solely on individual ratings. However, there is no reason by multiple ratings of the documents cannot be gathered.

Responses to the question about HOW a peer evaluation is conducted can also raise concerns about the validity and reliability of a method. Is training provided to the peer observers? Are the classroom observations announced or unannounced ahead of

time to the faculty member being evaluated? How is a consensus drawn in a documentation review? Must there be 100% agreement or true consensus or is it sufficient for a majority of the faculty to judge the documents satisfactory?

NEED FOR MORE WORK

In their article on "Peer evaluations for salary increases and promotions among college and university faculty members," Spaights and Bridges (1986) summarized some of the problems inherent in peer evaluation. These problem include: "judges who lack sufficient knowledge in the candidate's field, too few judges, close collegial relationship, adversarial relationships, lack of anonymity of the judges, and the overall problems inherent in a system that allows decisions to be made without a reasonable rationale". (p.405) These problems cut across the three methods of peer evaluation discussed in this paper. Yet, we believe that the earlier discussion of the why, who, what, when, and hows of peer evaluation revealed a strong case for further use of and research conducted with the peer documentation review technique.

Documentation review should not have to be "sold" to faculty who are somewhat comfortable with the procedures used in review manuscripts for publications, papers for conference presentation, and portfolios for promotion and tenure relative to research and service. Because of the traditions of document review, there should exist an adequate number of experience faculty to sit on teaching review committees. The method doesn't require that faculty place themselves in an awkward, unfamiliar role as classroom observer that requires considerable time and resources. Peer documentation avoids any

controversy regarding the evaluation of style by focusing, instead, on written documents. The method is based on evidence in hand rather than on general impressions in the minds of friends and foes. However, we are not naive enough to think that there will be no bias in the ratings. Judgments of the type of examinations used and the rigor in the grading of student papers and products are likely to require some systematic guidelines.

There is tremendous interest today in evaluating teaching in higher education. The newest evaluation fad is the development and evaluation of teaching portfolios. We are told that Seldin's (1991) booklet, "The Teaching Portfolio," is becoming a best seller. The booklet is well written and should be quite useful to administrators and faulty interested in peer documentation review. However, his portfolio contents do not differ much from what many institutions have been requiring in P & T documentation. Peer documentation review is not new.

Whereas peer documentation is not new it is not the methodology that first come to mind when someone asks for peer evaluation. Peer evaluation is often juxtapositioned with student evaluation or observation of the classroom behaviors of a professor. Peer observation is the first tried with its many problems and costs and because of these problems and costs many institutions choose not to institutionalize a peer evaluation process except at the final tenure and promotion decision. It is important for institutions to supplement student rating information with peer evaluations in annual reviews. For this reason we need to support and study the conduct of periodic or annual documentation reviews.

Researchers interested in peer evaluations should consider replicating some of the

earlier studies that compared peer ratings with student ratings, or peer ratings with measures of student learning. However, this round of studies should collect peer ratings through documentation reviews rather than through observations or general impressions.

In addition to conducting more research with peer documentation reviews, we need to do more to study the evaluation process. Just the other day a colleague at another university called and asked for some advice on how to evaluate teaching portfolios. After a small discussion we concluded that there has been much written about what goes into a portfolio but very little on how to evaluate one once it is completed.

Most institutions of higher education have developed faculty evaluation systems which are built either by tradition or contract. Unfortunately, many of these systems do not include a systematic peer evaluation procedure (Hazlett, 1990; Seldin, 1984) Yet, we know that a valid and reliable faculty evaluation system should incorporate different types of information collected from various sources. (Braskamp, Brandenburg, and Ory, 1984) Peer evaluation should be included in a comprehensive system. To fail to add a peer evaluation component because of our beliefs about peer observations would be a mistake.

REFERENCES

Batista, E.E. (1976) The place of colleague evaluation in the appraisal of college teaching. Research in Higher Education, 4, 257-271.

Blackburn, R. T. and Clark, M. J. (1975) An assessment of faculty performance: Some correlates between administrator, colleague, student, and self-ratings. Sociology of Education, 48, 252-256.

Braskamp, L., Brandenburg, D. and Ory, J. (1984) Evaluating teaching effectiveness. Beverly Hills, CA: Sage.

Cahn, S. M. (1986) Saints and Scamps Ethics in Academia. Totowa, NJ: Rowman & Littlefield.

Centra, J. (1975) Colleagues as raters of classroom instruction. Journal of Higher Education, 46, 327-337.

Doyle, K. O. and Crichton, L. I. (1978) Student, peer, and self-evaluations of college instructors. Journal of Educational Psychology, 70, 815-826.

Feldman, K. A. (1987) Research productivity and teaching effectiveness. Research in Higher Education. 26, 227-291.

French-Lazovick, G. (1980) Peer Review: Documentary Evidence in the Evaluation of Teaching. In J. Millman (Ed.) Handbook of teaching evaluation. Beverly Hills, CA: Sage.

Haziett, D. Evaluation of teaching effectiveness in the Association of American Universities (AAU) institutions: A Survey of current policies and practices, (unpublished manuscript).

Grassmuck, K. (1990) Some Research Universities contemplate Sweeping changes, ranging from Management and Tenure to Teaching Methods. The Chronicle of Higher Education. 37(2), p. A1 A29-A31.

Kremer, J. (1990) Construct validity of multiple measures of teaching, research, and service and reliability of peer ratings. Journal of Educational Psychology, 82, 213-218.

Kremer, J. (1991) Identifying Faculty Types Using Peer Ratings of Teaching, Research, and Service. Research in Higher Education, 32, 351-362.

Medley, D. M. and Mitzel, H. E. (1963) Measuring classroom behavior by systematic

observation, in N. Gage (ed.) Handbook of Research on Teaching. Chicago: Rand McNally.

Root, L. S. (1987) Faculty evaluation: Reliability of peer assessments of research, teaching and service. Research in Higher Education, 26, 71-84.

Rosenshine, B. and Furst, N. (1973) The use of direct observation study teaching, in R. Travers (ed.) Second Handbook of Research on Teaching. Chicago: Rand McNally.

Scriven, M. (1987) Duty-based teacher evaluation. Journal of Personnel Evaluation in Education, 1, 9-23.

Scriven, M. (September, 1991) Report on the teacher evaluation models project. Newsletter for the Center for Research on Educational Accountability and Teacher Evaluation (CREATE).

Seldin, P. (1984) Changing practices in faculty evaluation. San Francisco: Jossey Bass.

Seldin, P. (1991) The teaching portfolio. Boston: Anker Publishing.

Spaights, E. and Bridges, E. (1986) Peer evaluations for salary increases and promotions among college and university faculty members. North Central Association Quarterly, 60, 403-410.